SYNOPSIS

ON

A HYBRID APPROACH FOR INFORMATION RETRIEVAL USING BIG DATA ANALYTICS FOR MULTIMEDIA DATA

SUBMITTED TO

SHRI RAMDEOBABA COLLEGE OF ENGINEERING AND MANAGEMENT, NAGPUR

FOR THE DEGREE

OF

DOCTOR OF PHILOSOPHY

IN

(COMPUTER SCIENCE)

By

KHUSHBOO KHURANA

UNDER THE GUIDANCE OF

DR. M. B. CHANDAK

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SHRI RAMDEOBABA COLLEGE OF ENGINEERING AND MANAGEMENT, NAGPUR, MAHARASHTRA - 440015
Year- 2016
RESEARCH TITLE:
A Hybrid Approach for Information Retrieval using Big Data Analytics for Multimedia Data

1. ABSTRACT

The amount of data in the world is exploding. Data is being collected and stored at unprecedented rates. The challenge is not only to store and manage the vast volume of data, but also to analyze and extract meaningful value from it. Moreover, the unstructured type of data is has increased due to advancement in the technology. Unstructured data refers to information that either does not have a pre-defined data model or does not fit well into relational tables. Unstructured data is the fastest growing type of data, some example could be imagery, sensors, telemetry, video, documents, log files, and email data files. There are several techniques to address this problem space of unstructured analytics.

Multimedia data is prominent unstructured big data spread all over the Internet. It requires high processing power to extract useful information. Fast processing of multimedia data such as video is essential for criminal investigations, surveillance, news analysis, sports domain, emotion extraction, etc.

Analysis of multimedia data in reduced time is the latest research area. So, we have proposed a technique for analysis of multimedia data to extract meaningful information hidden in the big data.

2. INTRODUCTION

2.1 Big data

Advances in the technology has lead to increase in the amount of data. Analysis of this data is important to extract value from the abundance of data available. According to a 2011 IBM report, IBM Big Data Success Stories, every minute, 208,300 photos are uploaded to Facebook and 350,000 updates sent on Twitter. The rate at which the data is growing is very high. Businesses can use this data to understand whether customers are speaking about their companies in a positive or negative way, for example.

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data [1].

Big data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high-velocity capture, discovery, and/or analysis [2]. This definition delineates the four salient features of big data, i.e., volume, variety, velocity and value. Volume means the amount of data that needs to be handled is very large, and exabytes, zettabytes, and even higher amounts of data are described in big data applications. Variety means that the data is varied in nature, and structured, unstructured or semi-structured data needs to be properly combined to make the most of the analysis. Velocity means that a high rate of sampling is common in big data problems. Value means high yield will be achieved if the big data is used reasonable by analyzing correctly and accurately. Considering the velocity characteristic, algorithms which can improve efficiency of knowledge discovery are needed.

In 2011, Mckinsey's report [3] defined big data as, “datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.” This definition is subjective and does not define big data in terms of any particular metric. However, it tells which dataset can be considered as big data.
Big data analytics tends to focus on unstructured data, such as emails, videos, photos, and even posts on social media networks. Data mining is the field that enables computers to learn from data how to do predictions, diagnosis, recognition, using very little human intervention. Applications include medical diagnosis, predictions, user profiling, recommendations, image/speech recognitions.

Big data analytics can be categorized into two categories depending on the type of data for processing [4]:

1. Stream Processing: It is assumed that the data is fresh. The stream processing paradigm analyzes the data as soon as possible to derive its results. There is a continuous arrival of the data, it is fast and carries enormous volume only a small portion of the stream is stored in limited memory. One or few passes over the stream are made to find approximation results. Streaming processing theory and technology have been studied for decades. Representative open source systems include Storm, S4, and Kafka. The streaming processing paradigm is used for online applications, commonly at the second, or even millisecond, level.

2. Batch Processing: In the batch-processing paradigm, data are first stored and then analyzed. MapReduce [5] has become the dominant batch-processing model. The core idea of MapReduce is that data are first divided into small chunks. Next, these chunks are processed in parallel and in a distributed manner to generate intermediate results. The final result is derived by aggregating all the intermediate results.

After the big data analysis is done the result can be visualized. Visualization of big data is also an upcoming field of research.

2.2 Unstructured multimedia data analysis

Main challenges in big data include analysis, capture, curation, searching, sharing, storage, transfer, visualization and privacy violations. Big data is also comprehended as a moving target, where the information is bound to arrive and is bound to move forward for use. This term big data is mainly associated with larger sets of unstructured data. Unstructured data mainly includes text and multimedia content. Multimedia content analysis and storage is the highlight of research under consideration.

Video content is the growing market for both big data analysis and for business intelligence. The business analytics are using video content and information to generate new specifics for marketing and sales. The research is however mainly focused on analyzing the videos and how this unstructured data of videos can be used for managing as well as storing videos in an efficient and cost effective manner [6].

There are several techniques to address this problem space of unstructured analytics. The techniques share a common characteristics of scale-out, elasticity and high availability. MapReduce, in conjunction with the Hadoop Distributed File System (HDFS) and HBase database, as part of the Apache Hadoop project is a modern approach to analyze unstructured data. Hadoop clusters are an effective means of processing massive volumes of data, and can be improved with the right architectural approach [7].

3. PROBLEM DEFINITION

Focusing on big multimedia data understanding, the research aims to analyze multimedia data by feature extraction and mining this big data of feature set, to extract meaningful information.
4. LITERATURE REVIEW

Analysis of videos for information extraction is becoming an important research area. Due to the unstructured huge quantity of multimedia content it is termed as big data. Major challenges with video data are as follows:
- Unstructured
- Uncertainty
- High Volume

Design and implementation of a system to overcome such challenges is the need of today. To extract information from the big multimedia data various data mining tasks are available like clustering, outlier detection, structure mining and frequent pattern mining. Clustering refers to grouping to similar items. Outliers are the instances in the dataset that are distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error, hence they have to be excluded from the data set. An efficient distance based outlier detection algorithm has been proposed in [8]. Pattern mining is the discovery of interesting knowledge in the form of frequent occurring sets in the dataset.

4.1 Video data Analysis in Big data Environment

In [9], an extensible video processing framework in Apache Hadoop to parallelize video processing tasks in a cloud environment is presented. Except for video transcoding systems, there have been few systems that can perform various video processing in cloud computing environments. The framework employs FFmpeg for a video coder, and OpenCV for a image processing engine. To optimize the performance, it exploits MapReduce implementation details to minimize video image copy. Moreover, FFmpeg source code was modified and extended, to access and exchange essential data and information with Hadoop, effectively. A face tracking system was implemented on top of the framework for the demo, which traces the continuous face movements in a sequence of video frames.

An approach to the automated analysis is to split the problem into from video to metadata and metadata are used for interpretation tasks. In video analysis, content extraction is done and metadata is extracted from the video signal to form a “richly annotated temporal log” that contains symbolic representations of the activities, events, relationships, and other important content in the video [10]. Color feature based video content extraction is proposed in [11].

In [12], a novel system is proposed for 3D facial video analytic, which incorporates two-step facial shape motion retrieval method into map-reduce framework on cloud platform from videos.

Multimedia is Big Data, not only in terms of their volume, but also with respect to their heterogeneous nature. Manually annotating every single object in a large collection is not feasible. Therefore, content-based multimedia retrieval using sample objects as query input is increasingly becoming an important requirement for dealing with the data deluge. In [13], a novel multimedia retrieval system called as ADAM, is introduced. It is tailored to large collections and that is able to support both Boolean retrieval for structured data and similarity-based retrieval for feature vectors extracted from the multimedia objects is introduced. For efficient query processing in such big multimedia data, ADAM allows the distribution of the indexed collection to multiple shards and performs queries in a MapReduce style. Furthermore, it supports a signature-based indexing strategy for similarity search that heavily reduces the query time.

Metadata extraction and correction system for videos is presented in [14]. Firstly, based on structure of metadata attribute to be extracted is determined. A three-phase method to extract metadata. Secondly, a graph-based metadata correction approach for compensating
the accuracy of metadata extraction method is used. It fuses the big metadata of whole camera network, automatically detects suspicious metadata and corrects them based on the metadata spatial-temporal relationship and the image similarity. The system is implemented in a distributed fashion using Hadoop and HBase.

It is worthwhile to incorporate human knowledge with conventional machine learning approaches for big data analytics. In [15], a formal scenario recognition framework is presented where knowledge-based logic representation and reasoning is combined with data-based learning approach to enhance scenario recognition capabilities. This is achieved via multi-layered (hierarchical) processing. This approach constructs the hierarchical representation structure based on the semantic understanding of considered scenario, and transforms the structure into logic formulas. Conventional computer vision methods are applied for low-level events classification, and then logic based uncertainty reasoning is applied to determine scene content.

4.2 Feature extraction from Video Data

Video is an unstructured big multimedia data. It can be divided into scenes, shots and frames. Some of the features can be extracted at the frame level whereas some features need the complete video. Video can be characterized to have high-level and low-level features.

Low-level features can be extracted by frame processing. Color, shape, texture are some of the features. Spectral and spatial domains are the main methods used for feature separation. Motions of an object are studied from study of multiple images, separated by varying periods of time. Texture of an image is quantitatively described by its coarseness. The coarseness index is related to the spatial repetition period of the local structure. This feature is a distinguishing characteristic of an image [16].

Among other approaches, bag-of-features (BOF) methods have gained popularity due to their simplicity, wide range of application and high recognition accuracy. BOF methods represent actions by collections of local space-time descriptors aggregated over the video. Several alternative local descriptors have been proposed including histograms of flow orientations (HOF), motion boundary histograms (MBH), shapes of point trajectories, local trinary patterns and others [17].

Recent evaluation [18] demonstrates that MBH, HOF and HOG descriptors sampled along dense point trajectories outperform other methods. More recent extensions demonstrate improvements using motion stabilization and person trajectories. SIFT, Histogram of visual words, etc can also be used.

Matlab is very suitable for the analysis of features from the multimedia data.

5. PROPOSED METHODOLOGY

Multimedia data is characterized as unstructured and have high volume. This big data has to be analyzed for important information retrieval. The Multimedia data can be processed by metadata extraction and then further processing the metadata for information retrieval.

Analysis of multimedia data is the objective of the research. After feature extraction, the feature set will be analyzed to extract meaningful information.

The main aims of the research work are as under:

- Feature extraction by processing the unstructured multimedia data.
- Mining the Big data set to extract information.
- Using Parallel and distributed computing to handle Big data.
The figure shows the overview of the proposed system. Features extraction methodology will be designed for multimedia input data. The features will be analyzed to extract meaningful information from the input. It can be analyzed by various methods like clustering, pattern mining, etc. to extract information. A parallel and distributed technique will be designed for feature analysis and information retrieval.

REFERENCES

[8] Efficient distance based outlier detection on uncertain datasets of Gaussian distribution.